DOCUMENT RESUME

ED 138 623 TM 006 246

AUTHOR Lowry, Stephen R.
TITLE The Effect of Luck and Misinformation on the Discrepancy between Multiple-Choice Test Scores and True Ability.
PUB DATE [Apr 77]
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Ability; College Students; Guessing (Tests); *Multiple Choice Tests; Raw Scores; Response Style (Tests); Scoring Formulas; Statistical Analysis; Test Interpretation; *True Scores

ABSTRACT

The effects of luck and misinformation on ability of multiple-choice test scores to estimate examinee ability were investigated. Two measures of examinee ability were defined. Misinformation was shown to have little effect on ability of raw scores and a substantial effect on ability of corrected-for-guessing scores to estimate examinee ability. Luck was shown to be influential on ability of raw and corrected-for-guessing scores to estimate ability, though this influence was more pronounced for raw scores. Analysis of data confirmed all theoretical results. Furthermore, empirical results showed that raw scores were better estimators of examinee ability. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138623

THE EFFECT OF LUCK AND MISINFORMATION ON THE DISCREPANCY BETWEEN MULTIPLE-CHOICE TEST SCORES AND TRUE ABILITY

Stephen R. Lowry
University of Nebraska

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the Annual Meeting of
the American Educational Research Association
(New York, New York, April 1977)
(Session D-72, Small Roundtable, Division D)

TM006 246

The Effect of Luck and Misinformation on the Discrepancy Between Multiple-Choice Test Scores and True Ability

Stephen R. Lowry
University of Nebraska

Each examinee who took a multiple-choice test was assumed to have a fixed "true" performance level. Within th' ntext, various scoring procedures based on responses to multiple-choice t items, provided differing estimates of each examinee's performance level on the construct represented by the test items. The present study investigated the independent effects of luck and misinformation o the measurement of the ability of multiple-choice test scores to estimate examinee performance. Secondly, this study investigated whic of the two most prominent multiple-choice scoring methods was the better estimator of this examinee performance.

Definitions

For the most part, only two scoring methods are widely used in practical situations under the conventional response mode in which the examinees were asked to mark the correct (or best) choice from among the alternatives. They are as follows :

Raw Score - Score obtained as the sum of the item scores in which case an item is scored one if the correct choice was selected and zero otherwise.

Corrected-for-guessing score - Score S computed by the following formula: S=R-W/3 (four-choice items) in which R is the number right and W is the number wrong.

For purposes of this study, the preceeding test scores were analyzed only when the correction for guessing was applied with directions to guess on an item only when one or more choices could be eliminated and the raw scoring procedure was applied with directions to mark each item even if the mark reflected a pure guess. It is noted further, that three corrected-for-guessing scores were computed

empirically based on differences in the extent to which examinees responded to the admonition not to guess unless one or more choices could be eliminated prior to guessing.

Two measures of examinee performance, as estimated by the above multiple-choice test scores, were defined.

True Ability - Score composed of the sum of the number of items from which an examinee could confidently eliminate all wrong choices plus his expected gain from guessing on items from which he could eliminate at least one wrong choice leaving the correct choice.

Legitimate Ability - Examinee's actual number correct on all items from which he could eliminate at least one choice leaving the correct choice.

The criteria for goodness of an estimator is the ability to lead to the smallest absolute discrepancy between estimate and true or legitimate ability. The definitions of true and legitimate ability were consistent with classical true score theory in that misinformation had the same effect as the absence of information, both yielding zero score components.

Misinformation and luck, the effects which were investigated, were defined.

Misinformation Number of items on which the examinee confidently eliminated the correct choice as incorrect.

Luck The difference between expected guessing success and actual guessing success on items not involving misinformation. Luck was computed separately for items from which at least one wrong choice was eliminated and those on which the examinee guessed among all choices.

Theoretical Results

Various results were derived theoretically. Misinformation was shown to have little or no effect on the ability of multiple-choice test scores to estimate true or legitimate ability scores. In contrast, misinformation was

shown to have a substantial effect on the ability of corrected-for-quessing scores to estimate legitimate or true ability scores with the effect more pronounced on tests with fewer choices per item. Luck was shown to be influential on the ability of both raw scores and corrected-for-quessing scores to estimate true or legitimate ability scores, although this influence was more pronounced for raw scores.

Data Collection

To provide empirical verification of these results, data were collected from three sections of a college biology laboratory course with 47-86 students per section. All data reported herein reflected responses to two progress tests prior to the final examination. KR20 Reliability estimates ranged from .56 to .83. Examinees were required to respond in two distinct ways. Initially, they responded by marking the choices they confidently believed wrong receiving as scores the number of incorrect choices minus a severe penalty for each correct choice inadvertently marked as incorrect. Grades were assigned on the basis of these scores so that there was substantial motivation not to mark choices without knowledge. From these responses, true and legitimate ability scores and misinformation scores were computed. Finally, students were instructed to mark their best quess as to the correct choice for each item. These responses yielded the raw scores, the various luck scores, and, after deleting items based on differences in the extent to which examinees respond to the admonition not to quess, the three corrected-for-guessing scores.

The Relationship of Luck and Misinformation to the Discrepancy Between True or Legitimate Ability and Multiple-Choice Test Scores

To better assess the functional relationship between luck and misinformation and the discrepancy between true or legitimate ability and test scores, regression analysis will be used. Consider the following model.

$$\underline{Y} = \underline{X}\underline{b} + \underline{E}$$

Y, the dependent variable, refers to the vector of observed discrepancies between raw or corrected scores and true or legitimate ability. $\underset{\sim}{X}$ is the matrix of misinformation or luck scores, $\underline{B}$ is the vector of regression coefficients and $\underline{E}$ is the vector of errors distributed $N(\underline{0}, \sigma^2 I)$. A rejection of the null hypothesis is interpreted to mean that there is a significant regression effect of discrepancy on misinformation or luck.

Empirical Results

Misinformation

Table 1 contains the F statistics, observed α values which the data attained, and the calculated R^2 for the test of the regression effect of the discrepancies on misinformation for all six tests (Hypothesis 1).

Discrepancies 1,2,3,4 refer to the discrepancies between raw scores (D1) or the three forms of corrected scores (D2,D3, and D4) and legitimate ability. Discrepancies 5,6,7, and 8 refer to the discrepancies between raw scores (D5) or the three forms of corrected scores (D6,D7, and D8) and true ability.

The F statistics for D1 led to a rejection of the hypothesis ($p \leq .05$) and the conclusion that there was a significant regression effect of D1 on misinformation in only three of six tests ($R^2 \leq .14$). On the other hand, regression of D_2, D_3, and D_4 on misinformation was found to be significant in all six tests, in which cases, misinformation accounted for 21.5-68.7% of the variation of D_2, D_3, and D_4. While F values indicated a significant regression effect of D_5 on misinformation in four of the six tests, misinformation only accounted for a maximum of 24.5% of the variation in D_5. The effect of misinformation was found to be significant in all six tests for D_6, D_7, and D_8.

Thus, in confirmation of theory, misinformation was found to have little effect on the ability of raw scores to estimate legitimate or true ability. Furthermore, misinformation was found to have a substantial effect on the ability of corrected scores to estimate legitimate or true ability.

Luck

Table 2 contains the F statistics, the observed α values which the data attained, and the calculated R^2 for the test of the regression effect of the discrepancies on luck for all six tests (Hypothesis 2)

The F statistics indicated a significant effect of D_1 on luck for all six tests ($p \leq .05$) indicating that luck had an effect on the ability of raw scores to estimate legitimate ability. R^2 ranged from .106 to .656. A significant regression effect of D_2, D_3, and D_4 on luck was found in four of six, one of six, and four of six tests respectively. Luck accounted for a maximum of 32.7% of the variation in D_2, D_3, and D_4. The regression effect of D_5 on luck was found significant in all six tests with R^2 ranging from .856 - .930. The regression effects of D_6, D_7, and D_8 were also found to be significant for all six tests with R^2 ranging from .500 - .876. Thus, in alignment with the theoretical results, luck was found to be influential on the ability of raw scores and, to a lesser but substantial extent, corrected scores to estimate legitimate and true ability.

Discussion

It seems unreasonable to penalize an examinee for misinformation since an examinee with either misinformation or no information is unable to correctly answer a multiple-choice test item. On the other hand, if luck is assumed to be random, an examinee who took many tests would tend to have good luck only slightly more or less often than bad luck. Thus, even if the presence of different levels of luck was found to be influential on the ability of test scores to estimate legitimate or true ability, the average effect due to luck should be minimal since average luck over many tests should be small for each examinee.

Raw Scores and Corrected-for-Guessing Scores as Estimators of True and Legitimate Ability

In order to operationalize the comparison of the goodness of the corrected scores and raw scores as estimators of legitimate and true ability, a repeated

measurements test was conducted to test the hypothesis that the four absolute discrepancy means were equal. The errors for four discrepancies for each of N examinees were assumed to be distributed multivariate normal with mean $\underline{0}$ and variance $\underline{\Sigma}$.

Confidence intervals on the three contrasts between the discrepancies involving raw scores pairwise with each of the three corrected scores were computed.

Inasmuch as the above contrasts were the only ones of importance, inference will only be made from the tests of the hypothesis involving these contrasts. The test of each contrast involves testing the hypothesis that each of the two test scores in the contrast have equal mean absolute discrepancies. The hypothesis is rejected ($p \leq .05$) when the confidence interval does not include the zero point with the interpretation that the test score with the smaller absolute mean discrepancy is the better estimator of legitimate or true ability. Table 3 contains the means, standard deviations, confidence intervals, and F statistics for Hypothesis 3 (Legitimate Ability) and 4 (True Ability).

Legitimate Ability

All the contrasts for test 1 and 2 did not contain the zero point. Two contrasts for test 3 contained the zero point by a slight margin while the third contrast does not contain zero. For the last three tests, all contrasts did contain the zero point. However, while D_1 had slightly larger mean than D_2, D_3, or D_4 for test 6, it had the smallest mean for test 4 and 5. Consequently, since D_1 had the smallest mean discrepancy in 5 of 6 tests and the differences were significantly smaller in three of six tests, raw scores were taken to be the better estimates of legitimate ability.

True Ability

In comparing D_5, D_6, D_7, and D_8, it was determined that all three confidence intervals for test 1 did not contain the zero point, and since D_1 had the

smallest mean discrepancy, raw scores were taken to be the better estimators of true ability for test 1. All three contrasts for each of the remaining five tests contained the zero point. Furthermore D_1 had slightly larger mean absolute discrepancies in four tests.

Thus, the hypothesis of equality of mean absolute discrepancies was rejected for only one test. For that test, D_1 had the smaller mean discrepancy. It was further noted that four of the six tests exhibited larger mean absolute discrepancies for D_1 than D_2, D_3, and D_4. Therefore, there was insufficient evidence to conclude that raw scores or corrected scores were better estimates of true ability. To the extent that the data in the present study conformed to Lord's (1975) assumption that the only difference in answer sheets obtained under corrected-for-quessing instructions and raw scoring instructions is that the omitted responses in the former case are replaced by random quesses in the latter; the present results did not support Lord's conclusion that corrected-for-quessing scores are better estimators of an examinee's standing on a trait measured.

Conclusion

On the basis of the preceding discussion and results, the raw scoring procedure was recommended as the better estimator of legitimate ability for the following reasons.

1. Luck had more influence on the ability of raw scores to estimate legitimate ability while misinformation had more influence on the ability of corrected-for-quessing scores to estimate legitimate ability.

2. Raw scores were shown generally to have smaller mean absolute discrepancies between estimator and legitimate ability.

3. The raw scoring procedure was less complex both in terms of testing directions and score computations and thus more efficient.

While there was insufficient evidence to conclude that either raw scores or corrected-for-quessing scores were better estimators of true ability, the raw scoring procedure was recommended as the preferred estimator for the same reasons 1 and 3 above.

These recommendations were made subject to the following conditions.

1. Misinformation had no severe negative ramifications as would be the case with a surgeon in an operating room.

2. The instructors were interested in the scores that were the better estimators of an absolute criteria such as true or legitimate ability rather than a ranking of examinee scores which could be made based on correlations between the observed test scores and the ability scores of interest.

REFERENCE

Lord, F. M. Formula scoring and number-right scoring. *Journal of Educational Measurement* 1975, *12*, 7-11.

Table 1: F Statistics, Observed α Values and R^2 for the Test of the Regression Effect of the Discrepancies on Misinformation for Six Tests (Hypothesis 1)

F Statistics

	Tests	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_9
1	(N=86)	7.09	113.33	85.07	97.39	24.78	49.08	55.64	53.30
2	(N=86)	5.42	69.31	43.67	38.35	5.08	17.11	19.72	20.30
3	(N=47)	1.08	64.81	16.72	20.08	2.88	13.14	12.56	10.41
4	(N=47)	6.68	87.65	66.75	22.95	9.43	22.95	30.88	18.40
5	(N=78)	2.04	67.19	18.29	29.13	8.44	36.40	31.65	25.68
6	(N=78)	1.25	92.22	40.68	19.76	2.81	21.48	28.01	15.91
					Prob > F				
1		.0094	.0001	.0001	.0001	.0001	.0001	.0001	.0001
2		.0225	.0001	.0001	.0001	.0270	.0001	.0001	.0001
3		.3045	.0001	.0002	.0001	.0974	.0003	.0010	.0025
4		.0135	.0001	.0001	.0001	.0038	.0001	.0001	.0001
5		.157	.0001	.0001	.0001	.0050	.0001	.0001	.0001
6		.267	.0001	.0001	.0001	.0982	.0001	.0001	.0002
					R^2				
1		.0854	.5986	.5281	.5617	.2459	.3924	.4227	.4122
2		.0650	.4705	.3589	.3296	.0611	.1799	.2018	.2065
3		.0263	.6184	.2948	.3343	.0672	.2473	.2390	.2065
4		.1430	.6866	.6253	.3646	.1910	.3646	.4357	.3151
5		.0280	.4830	.2030	.2880	.1050	.3360	.3050	.2630
6		.0170	.5620	.3610	.2150	.0380	.2300	.2800	.1810

Table 2: F Statistics, Observed α Values and R^2 for the Test of the Regression Effect of the Discrepancies on Luck for Six Tests (Hypothesis 2)

F Statistics

		D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
1	(N=86)	9.02	36.89	11.22	24.32	1017.52	536.67	330.43	380.50
2	(N=86)	35.74	.104	3.41	.483	463.41	210.87	129.74	82.96
3	(N=47)	10.83	3.82	.552	5.79	263.39	223.88	104.71	126.64
4	(N=47)	76.10	18.75	1.53	11.94	463.87	225.73	112.87	79.24
5	(N=78)	61.65	16.41	.341	28.68	740.34	375.87	135.52	286.70
6	(N=78)	61.54	5.23	.555	1.77	567.88	262.18	88.76	72.07

Prob > F

1	.0036	.0001	.0013	.0001	.0001	.0001	.0001	.0001
2	.0001	.7479	.0684	.4890	.0001	.0001	.0001	.0001
3	.0021	.0577	.4620	.0210	.0001	.0001	.0001	.0001
4	.0001	.0001	.2230	.0013	.0001	.0001	.0001	.0001
5	.0001	.0001	.5610	.0001	.0001	.0001	.0001	.0001
6	.0001	.0250	.4590	.1870	.0001	.0001	.0001	.0001

R^2

1	.1061	.3268	.1286	.2424	.9305	.8759	.8130	.8335
2	.3142	.0013	.0419	.0062	.8559	.7300	.6245	.5154
3	.2131	.0872	.0136	.1265	.8682	.8484	.7236	.7600
4	.6555	.3190	.0368	.2299	.9206	.8495	.7380	.6646
5	.4613	.1860	.0050	.2850	.9114	.8390	.6530	.7990
6	.4608	.0680	.0080	.0240	.8875	.7850	.5520	.5000

Table 3: Means, Standard Deviations, Confidence Intervals, and F Statistics for Hypotheses Three and Four

$H_3: \mu_{D_1} = \mu_{D_2} = \mu_{D_3} = \mu_{D_4}$

		1	2	3	4	5	6
	1	.603	.438	1.10	1.59	1.24	1.62
Mean	2	2.86	1.33	1.91	2.10	1.58	1.18
Discrepancies	3	3.02	1.39	2.27	2.28	1.75	1.46
	4	2.81	1.38	1.67	2.35	1.43	1.36
	1	1.11	1.42	1.81	3.15	1.79	2.81
Standard	2	1.29	.656	.941	1.21	.912	1.06
Deviations	3	1.31	.718	1.07	1.16	1.14	1.40
	4	1.41	.822	1.09	1.47	1.04	1.61
	1,2 L	-2.97	-1.56	-2.04	-2.69	-1.18	- .742
	U	-1.55	- .229	.416	1.69	.500	1.63
Confidence	1,3 L	-3.09	-1.63	-2.42	-2.80	-1.31	-1.06
Intervals	U	-1.74	- .267	- .072	1.43	.305	1.38
	1,4 L	-3.03	-1.52	-1.92	-2.36	-1.12	- .672
	U	-1.38	- .357	.761	.849	.751	1.18
$F_{3, n-3}$		60.73	11.76	10.35	5.26	4.24	3.39

$H_4: \mu_{D_5} = \mu_{D_6} = \mu_{D_7} = \mu_{D_8}$

		1	2	3	4	5	6
	5	1.77	1.18	2.35	2.81	2.63	2.82
Mean	6	3.19	1.25	1.89	2.27	1.62	1.57
Discrepancies	7	3.21	1.22	1.91	2.13	1.59	1.54
	8	3.19	1.41	1.97	2.61	1.95	1.98
	5	1.94	1.60	2.61	3.68	2.38	3.13
Standard	6	2.24	.948	1.50	1.80	1.33	1.52
Deviations	7	2.14	.902	1.30	1.64	1.25	1.45
	8	2.22	1.05	1.84	2.31	1.43	2.01
	1,2 L	-2.58	- .863	-1.14	-1.79	- .101	- .097
	U	- .250	.721	2.04	2.87	2.12	2.59
Confidence	1,3 L	-2.59	- .834	-1.28	-1.74	- .103	- .109
Intervals	U	- .285	.759	2.16	3.09	2.17	2.68
	1,4 L	-2.57	- .887	-1.05	-1.60	- .282	- .198
	U	- .256	.421	1.80	1.98	1.62	1.88
$F_{3, n-3}$		6.20	3.03	.394	1.20	4.60	3.48